Journal of Toxicology and Environmental Health, Part A
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713667303

Evaluating Sufficient Similarity for Disinfection By-Product (DBP) Mixtures: Multivariate Statistical Procedures

Paul I. Feder [a]; Zhenxu J. Ma [a]; Richard J. Bull [b]; Linda K. Teuschler [c]; Kathleen M. Schenck [c]; Jane E. Simmons [d]; Glenn Rice [c]
[a] Battelle, Statistics and Information Analysis, Columbus, Ohio, USA [b] MoBull Consulting, Richland, Washington, USA [c] U.S. Environmental Protection Agency, Cincinnati, Ohio, USA [d] National Health and Environmental Effects Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA

Online Publication Date: 01 January 2009

To cite this Article Feder, Paul I., Ma, Zhenxu J., Bull, Richard J., Teuschler, Linda K., Schenck, Kathleen M., Simmons, Jane E. and Rice, Glenn(2009)'Evaluating Sufficient Similarity for Disinfection By-Product (DBP) Mixtures: Multivariate Statistical Procedures',Journal of Toxicology and Environmental Health, Part A,72:7,468 — 481

To link to this Article: DOI: 10.1080/15287390802608965

URL: http://dx.doi.org/10.1080/15287390802608965

Journal of Toxicology and Environmental Health, Part A, 72: 468–481, 2009
This article is not subject to U.S. copyright
ISSN: 1528-7394 print / 1087-2620 online
DOI: 10.1080/15287390802608965

Evaluating Sufficient Similarity for Disinfection By-Product (DBP) Mixtures: Multivariate Statistical Procedures

Paul I. Feder[1], Zhenxu J. Ma[1], Richard J. Bull[2], Linda K. Teuschler[3], Kathleen M. Schenck[3], Jane E. Simmons[4], and Glenn Rice[3]

[1]*Battelle, Statistics and Information Analysis, Columbus, Ohio,* [2]*MoBull Consulting, Richland, Washington,* [3]*U.S. Environmental Protection Agency, Cincinnati, Ohio, and* [4]*National Health and Environmental Effects Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA*

For evaluation of the adverse health effects associated with exposures to complex chemical mixtures in the environment, the U.S. Environmental Protection Agency (EPA) (2000) states, “if no data are available on the mixture of concern, but health effects data are available on a similar mixture . . . a decision must be made whether the mixture on which health effects are available is ‘sufficiently’ similar to the mixture of concern to permit a risk assessment.” This article provides a detailed discussion of statistical considerations for evaluation of the similarity of mixtures. Multivariate statistical procedures are suggested to determine whether individual samples of drinking-water disinfection by-products (DBPs) vary significantly from a group of samples that are considered to be similar. The application of principal components analysis to (1) reduce the dimensionality of the vectors of water samples and (2) permit visualization and statistical comparisons in lower dimensional space is suggested. Formal analysis of variance tests of homogeneity are illustrated. These multivariate statistical procedures are applied to a data set describing samples from multiple water treatment plants. Essential data required for carrying out sensitive analyses include (1) identification and measurement of toxicologically sensitive process input and output characteristics, and (2) estimates of variability within the data to construct statistically efficient estimates and tests.

Many contaminants occur in the environment as complex mixtures, which are generally comprised of many, perhaps hundreds of, chemical components. The proportions of these components vary depending on how the mixture was formed and the fate of the mixture in the environment. Furthermore, a portion of the complex mixture may be poorly characterized chemically, with some components unknown.

The U.S. EPA has developed guidance that discusses concepts and suggested procedures for the health risk assessment of exposures to multiple chemicals, including complex mixtures (U.S. EPA, 1986, 2000). Ideally, environmental health risk assessments are conducted using dose-response data developed from the mixture of concern. However when dose-response data are not available for the chemical mixture to which individuals are exposed (i.e., the mixture of concern), U.S. EPA (2000) recommends the use of dose-response data from a *similar* mixture (i.e., the tested mixture). Consequently, the determination of whether the mixture of concern is “sufficiently similar” to a tested mixture or a group of tested mixtures becomes central to the use of whole mixture methods. The U.S. EPA guidance states that differences in components or the components proportions be considered in analyses of similarity and that the determination of “sufficient similarity” needs to consider the uncertainties associated with using data on the tested mixture as a surrogate for another environmental mixture. The guidance does not, however, provide specific direction on methods with which to approach such a problem.

The chemical disinfection of drinking water results in the formation of complex mixtures comprised of a large and complex array of by-products. Bull et al. (2001, 2008a, 2008b) discuss in detail the various classes of disinfection by-products (DBP) that occur in drinking water and are associated with various adverse health consequences. They characterize input factors to the water treatment process that affect the formation of by-products, the variation in the output DBP associated with these factors, and some of the toxicological implications

The views expressed in this article are those of the individual authors and do not necessarily reflect the views and policies of the U.S. EPA. Those sections prepared by U.S. EPA scientists have been reviewed in accordance with U.S. EPA peer and administrative review policies and approved for presentation and publication. Mention of trade names or commercial products does not constitute endorsement or recommendations for use.

The authors acknowledge and appreciate the many helpful review comments and suggestions of Richard C. Hertzberg, PhD (Emory University), and David Farrar, PhD (U.S. EPA/ORD/NCEA). These comments greatly improved this article.

Address correspondence to Paul I. Feder, Battelle, Statistics and Information Analysis, 505 King Avenue, Columbus, OH 43201-2693, USA. E-mail: feder@battelle.org

of this variation. Bull et al. (2001, 2009a, 2009b) note that the composition of DBP mixtures in finished drinking water is a complex function of source water characteristics, treatment process characteristics, and modifications to physical and chemical characteristics of the finished water that occur during the distribution of the water to houses and industry (e.g., changes in pH or temperature, additional water treatment using chemical substances). This complex mixture varies across treatment plants located in different places, over time within the same treatment plant, and in the treated waters as they travel through the distribution system.

Rice et al. (2009) present an overview of proposed methods for assessing similarity of complex mixtures. Rice et al. (2009) discuss methods based on whole mixtures and methods based on individual mixture component data. They state that if whole-mixtures data are available then whole-mixtures methods are preferred because they are subject to less uncertainty. The statistical methods discussed in this article are based on whole-mixtures data.

The focus of this article is to (1) present and illustrate multivariate statistical analysis and graphical methods that assess the degree of similarity of multiple input water sources to treatment plants, using data from Schenck et al. (2009), and (2) assess the similarity of the finished water and the distribution system water emanating from those five treatment plants. The source water category and the chlorine dose used at each plant are provided in Table 1.

The statistical methods are illustrated based on selected characteristics of the input and output water that were considered to be most important in affecting the degree of similarity (Table 2). These data describe the water source, the treatment characteristics, and the finished water and distribution water characteristics. Many of these characteristics are "summary variables," which describe the whole mixture, rather than single chemicals within the mixture. Examples of such variables that describe the DBP mixture include total organic halogens (TOX), mutagenic activity, total trihalomethanes (TTHM), haloacetic acids (HHA), and percent TTHM and HHA that consists of brominated compounds. Use of summary variables is essential to evaluating similarity among mixtures because they account for

TABLE 1
The Five Treatment Plants and the Month, Year During Which They Were Monitored, the Water Sources, and the Chlorine Treatment Characteristics

Treatment plant #	Date	Water source	Chlorine dose(s)
1	September, 1995	Ground water	2.12 mg/L
2	March, 1996	Surface water Creek	4.1 mg/L
3	June, 1996	Surface water River	4.9 and 0.4 mg/L
4	September, 1997	Surface water Two lakes	1.92 and 1.98 mg/L
5	September, 1998	Surface water Pond	1.98 mg/L

TABLE 2
Input and Output Factors Affecting the Chlorination By-products in Drinking Water

Process	Characteristics
Source Water (Input)	• Surface water vs. ground water • Total Organic Carbon (TOC) • UV absorbance • Bromide level
Treatment Plant (Input)	• Form of chlorine input—Gas vs. liquid (hypochlorite solution) • Chlorine concentration (mg/L source water) added during chlorination step(s) • NH_3 added to treated water at or following chlorination step (chloramination) • pH of water at chlorination step • Temperature of water at chlorination step and beyond
Finished and distribution Water (Output).	• TOC (alternatively UV) • Total organic halogen (TOX) • Total trihalomethanes (THM) • Percent brominated • Percent mixed (brominated-chlorinated) • Percent chlorinated; • Six haloacetic acids (HAA6) • Percent brominated • Percent mixed (brominated-chlorinated) • Percent chlorinated • Total haloacetonitriles • Percent brominated • Percent mixed (brominated-chlorinated) • Percent chlorinated • Aldehydes (e.g., formaldehyde, acetaldehyde) • Haloketones • Miscellaneous halogenated DBPs • N-nitrosodimethylmine (NDMA, other nitrosamines) • Chlorate • Mutagenic activity

the unidentified chemical components, as well as for biological and chemical interactions among them.

For the input water characteristics, the "unit of analysis" is the water source–treatment plant process combination. For the output water characteristics, the "unit of analysis" is the individual water sample corresponding to a particular treatment plant and time/place in the system at which the sample was obtained. At each of the five treatment plants, one finished water sample and two distribution water samples were analyzed (i.e., 15 water samples).

CHARACTERIZATION OF SIMILARITY WITH RESPECT TO PROCESS INPUT VARIABLES

Nine important source water and treatment plant process input characteristics are displayed within the first two blocks of Table 2. These include two dichotomous (i.e., having two discrete levels) variables and seven measured variables. These nine input characteristics were selected from the entirety of input characteristics, based on scientific judgment concerning those input water and treatment plant characteristics that are thought to be most highly associated with the toxicity of the finished water.

Statistical inferences to assess "sufficient similarity" among water sources require:

- A reference set of input water source/treatment plant process combinations that are considered to be statistically homogeneous.
- Assessment of statistical variation among the water sources that are assumed to be homogeneous.
- Statistical inferences as to:
    - Whether any water samples in the reference set are outliers with respect to the homogeneous set.
    - Whether a new water sample comes from the same population as the reference set.
    - Whether distinct subsets of water samples are statistically distinguishable.

For the statistical analyses to be meaningful from a toxicological perspective, the process input characteristics (i.e., source water, treatment plant, and disinfection process variables) that are included in the statistical analyses need to be sufficient to characterize the family of similar mixtures based on the reference data set and to predict the classes of process output DBP that are important in characterizing the toxicity of the finished water and the distribution water. This reference distribution is then used as a yardstick against which to compare additional processes.

Although most of the input water characteristics shown in Table 2 are chemical characteristics, the similarity of mixtures with respect to toxicological effects is of principal interest from a public health perspective. Thus, similarity of chemical characteristics among mixtures is of interest primarily to the extent that the similarity of chemical characteristics implies the similarity of toxicological characteristics.

An important consideration in interpreting the results of the analyses discussed in this article is the "universe" of water sources and treatment processes to which the similarity inferences apply. Two different situations might apply.

- The treatment plants included in the analyses may constitute the "universe" of treatment plants about which inferences are to be made based on the data. For example, this might be the case if the treatment plants in the data set were the totality of plants utilizing a single water source (e.g., Lake Ontario) and inferences were to be restricted to treatment plants utilizing that water source.
- Alternatively, the treatment plants included in the analyses may be a random sample drawn from a larger "population" of treatment plants. Inferences are to be made about the larger population. In this situation the reference set of sampled treatment plants must be sufficiently broad that it is representative of the population.

The examples discussed in this article relate to the data presented in the Schenck et al. (2008). The five treatment plants in the Schenck et al. (2008) article are not a randomly representative selected sample from a larger population of treatment plants. Thus the statistical inferences based on these data pertain to comparisons among the particular treatment plants in the data set.

Among the nine input characteristics displayed in the first and second blocks of Table 2, the type of water (ground or surface) and the form of chlorine (gas or liquid) are each dichotomous (i.e., they have two discrete levels). Let *i* index the type of water source ($i = 1, 2$) and let *j* index the form of chlorination ($j = 1, 2$). The four combinations of *i* and *j* define four "strata," each representing a distinct joint distribution for the other *p* source water input factors and treatment plant input factors (here, $p = 7$). For each combination (i, j), a *p*-dimensional joint distribution exists with mean vector μ_{ij} and covariance matrix Σ_{ij}. Let X_{ij} denote the *p*-dimensional vector of input characteristics for a sample in stratum (i, j). Assume the possibly after taking preliminary transformations the random vector X_{ij} is approximately distributed according to a multivariate normal distribution with mean vector μ_{ij} and covariance matrix $\Sigma_{ij.}$ That is,

$$X_{ij} \sim MVN(\mu_{ij}, \Sigma_{ij}) \quad i, j = 1, 2$$

The *k*th element of the mean vector μ_{ij} is the mean of the *k*th response variable in stratum (i, j). The (h, k)th element of the covariance matrix Σ_{ij} is the covariance between the *h*th and *k*th response variables in stratum (i, j). The assumption of multivariate normality underlies the statistical tests for equality of water sources and the outlier detection procedures discussed later in the article, With sufficient data, the distribution parameters within each stratum can be estimated based on a linear model:

$$\mu_{ij} = \mu_0 + \alpha I_{Si} + \beta I_{Fj} + \gamma I_{Si} I_{Fj}$$
$$\Sigma_{ij} = \Sigma_0 + A\, I_{Si} + B\, I_{Fj} + G I_{Si} I_{Fj}$$

Here I_{Si} = 1 for surface water and I_{Si} = 0 for groundwater. I_{Fj} = 1 for liquid hypochlorite and I_{Fj} = 0 for gaseous chlorine. μ_0 and Σ_0 represent the (baseline) mean vector and covariance matrix for groundwater (i = 2) and gaseous chlorine (j = 2); α and **A** represent the difference in mean vector, covariance matrix between the surface water (i = 1) and the baseline groundwater (i = 2) with gaseous chlorine treatment; β and B represent the difference in mean vector, covariance matrix between liquid hypochlorite (j = 1) and the baseline gaseous chlorine (j = 2) with groundwater source; and γ and G represent the difference in effects between liquid hypochlorite and gaseous chlorine on the mean vector and covariance matrix, between surface water and groundwater (i.e., the "interactions" between water source and treatment plant).

In the most general situation, a separate distribution exists within each stratum. However, if α, β, γ, A, B, and G are all 0, this reduces to a common distribution among all the strata. If only γ and G are 0, then the interactions are 0 and a "main effects model" holds. This means that the effect of water source is the same irrespective of chlorine administration method and conversely the effect of chlorine administration method is the same irrespective of water source.

Within stratum (i, j), assume that N_{ij} observations $X_{ij1}, \ldots, X_{ijNij}$ exist. Calculate the sample mean and sample covariance matrix as

$$\bar{X}_{ij} = \frac{1}{N_{ij}} \sum_{k=1}^{N_{ij}} X_{ijk}$$
$$S_{ij} = \frac{1}{N_{ij}-1} \sum_{k=1}^{N_{ij}} (X_{ijk} - \bar{X}_{ij})(X_{ijk} - \bar{X}_{ij})'$$

Estimation of separate parameters within each stratum requires sufficient data for each combination of source water and chlorine administration. It is required that there be a minimum of p + 1 (= 8) distinct water samples within each stratum in order that the sample covariance matrix S_{ij} can be fully estimated so that there are no linear dependencies among variables and so that the inverse of S_{ij} can be determined.

In the absence of sufficient data within each stratum to estimate separate covariances or to verify different covariances among distributions, the models of underlying distributions may be simplified by pooling data across strata to estimate common parameters, if an assumption of common parameters across strata is reasonable. For example, if the variances and covariances among variables can be assumed to be equal across strata, then a pooled sample covariance matrix can be calculated as

$$S = \frac{1}{\sum_{i,j}(N_{ij}-1)} \sum_{i,j} (N_{ij}-1) S_{ij}$$

It is necessary that the degrees of freedom Σ_{ij} (N_{ij} − 1) be at least p (= 7) in order for the sample covariance matrix to be fully estimable (i.e., is of full rank) so that the inverse S^{-1} can be determined.

CHARACTERIZATION OF SIMILARITY WITH RESPECT TO PROCESS OUTPUT VARIABLES

The statistical inference questions for studying the similarity of source water and treatment plant input characteristics are also relevant for studying the similarity of process output characteristics for multiple water samples. In the Schenck et al. (2009) data set the process outputs are summarized by a vector of up to 20 chemical characteristics of the finished water and the distribution water, as shown in the third block of Table 2. The output data and the questions to be answered are richer and are more varied than for the process input data.

The physical and chemical characteristics of the process outputs were determined in the Schenck et al. (2008) data on the finished water as well as the downstream distribution samples. A natural inference question is to assess the similarity between the distribution water and the finished water. Each treatment plant provides water sample information from matched finished water samples and downstream distribution samples. This results in a two-sample matched-pairs analysis situation.

Because of the relatively small number of treatment plants included in the data, the vector of output characteristics was reduced to a subset of somewhat lower dimension in the examples, so that the inference procedures could be illustrated. Thus, the examples illustrate the types of multivariate comparison methods that can be carried out, but are not as extensive a multivariate comparison as one might choose to carry out if a more extensive set of treatment plants were available.

STATISTICAL METHODS FOR ANALYSES AND DISPLAYS

Statistical analysis procedures are discussed next, with various analysis objectives that are useful in understanding the structure of the data.

- Similarity of a new water sample compared to a set of reference samples.
- Two-sample comparison between the set of finished water samples and the set of distribution water samples by a two-sample analysis, not accounting for matching within treatment plants.
- Identification of outlying samples from within a group of possibly similar mixtures.

- Matched pairs comparison between the set of finished water samples and the set of distribution water samples by a two-way multivariate analysis of variance, accounting for matching within treatment plants.
- Presentation of graphical and inferential results in the reduced dimensionality space of principal components.

Comparison of a New Treatment Process to a Reference Set of Similar Processes

Consider first the simplest situation. Assume that a common reference set distribution exists across the (water source (*i*) and chlorination method (*j*)) strata, and that it can be modeled as normally distributed with mean vector μ and covariance matrix Σ, MVN(μ,Σ). Assume that $N \equiv \Sigma_{ij} N_{ij}$ independent water samples are selected from this single population of similar processes. Denote the *p*-dimensional response vector for the reference set as X_{ij}. Suppose that an additional treatment plant has a *p*-dimensional response vector *Y*, and it is desired to test the null hypothesis that the process that generates *Y* is similar to the process that generates the X_{ij}. Under the assumption that the distribution of *Y* has the same covariance matrix Σ as the common distribution of the X_{ij} terms, Hotelling's T^2 statistic (Morrison, 1976, p. 131) can be utilized to compare the distribution of *Y* with that of the *X*s:

$$T^2 = (Y - \bar{X})'[(1+\tfrac{1}{N})S]^{-1}(Y - \bar{X})$$

where $\bar{X} \equiv \frac{1}{N}\Sigma_{ij}\, N_{ij}\, \bar{X}_{ij}$ and *S* are as defined previously. Under the null hypothesis of equality of means the distribution of T^2 is proportional to that of a central *F* distribution with *p* and ν –(p – 1) degrees of freedom:

$$\tfrac{\nu-(p-1)}{p\nu} T^2 \sim F_{p,\nu-(p-1)}$$

Under an alternative hypothesis that the process *Y* differs from that of the X_{ij}'s, then $Y \sim$ MVN(μ + δ,Σ), where δ is the incremental change in the mean vector μ of water characteristics. In that case, the Hotelling's T^2 is distributed proportional to a noncentral *F* distribution with the noncentrality parameter λ related to δ, namely,

$$\frac{\nu-(p-1)}{p\nu} T^2 \sim F'_{\nu-(p-1)}(\lambda)$$

where

$$\lambda = (\delta)'\left[(1+\tfrac{1}{N})\Sigma\right]^{-1}(\delta)$$

Here λ is referred to as a "noncentrality parameter." It measures how far the "true" mean lies from the null hypothesis space (after normalizing by the variances and covariances among the variables). This specification of the alternative distribution permits the construction of contours of δ that correspond to various levels of power (e.g., 90%) to detect departures from the common distribution among the sample of treatment processes that are considered to be similar. When λ is 0, the noncentral *F* distribution reduces to the central *F* distribution.

The toxicological sensitivity of this simplest inference procedure to discriminate between similar and dissimilar processes on a statistical basis depends on whether the response values for processes that lead to dissimilar mixtures on toxicological grounds are sufficiently far removed from the reference set distribution so that λ is large enough to have the desired level of power.

This relatively simple procedure is adapted later to identify individual vectors of water characteristics that may be outliers from the overall populations.

Two-Sample Analysis (Unmatched)

Consider the simple scenario where the finished water samples can be regarded as a single sample and the distribution system water can be regarded as a second sample, with no pairing between samples. Let F_i, i =1, . . . , N_F denote the finished water samples, and let D_j, j = 1, . . . , N_D denote the distribution water samples. Assume that (possibly after taking transformations of the data) *F* has a multivariate normal distribution with mean and covariance matrix μ_F and Σ_F, and that *D* has a multivariate normal distribution with mean and covariance matrix μ_D and Σ_D. Assume that μ_D, Σ_D, μ_F, Σ_F are estimated by $\bar{D}, S_D, \bar{F}, S_F$ respectively. Note that with a *p*-dimensional response, (N_F – 1) and (N_D – 1) must each be at least *p* in order for the estimated covariance matrices S_D and S_F to be fully estimable (i.e., of full rank).

The null hypothesis to be tested is

$$H_0\!:\ \mu_F = \mu_D$$

versus the alternative hypothesis

$$H_1\!:\ \mu_F \neq \mu_D$$

Under the null hypothesis the finished water and distribution water mean responses are equal for each variable. Under the alternative hypothesis they differ for at least one variable.

If there are sufficient numbers of finished water samples and distribution water samples to estimate separate covarianc**e** matrices S_F and $S_{D,}$ a preliminary assessment of the equality of the covariance matrices Σ_F and Σ_D can be carried out. Morrison (1976, p. 252ff) presents a test of the equality of the covariance matrices due to Box (1949), based on the sample covariance matrices. The performance of the test is sensitive to the assumption of multivariate normality. If the Box test rejects H_0: $\Sigma_F = \Sigma_D$, it is necessary to use a heterogeneous variance T^2 statistic to compare μ_X and μ_Y. Such a statistic is a direct generalization of the separate variance *t* statistic for univariate comparisons.

For purposes of analysis of the Schenck et al. (2009) data there is not a sufficient number of water sources to estimate the covariance matrices for the finished water and for the distribution water separately, or to carry out a preliminary test. The homogeneity of covariances between the finished water and the distribution water is assumed and the covariance matrices are pooled. The pooled sample covariance matrix is

$$S = \frac{1}{(N_F - 1) + (N_D - 1)} [(N_F - 1) S_F + (N_D - 1) S_D]$$

With a *p*-dimensional response $(N_F - 1) + (N_D - 1)$ must be at least *p* in order for the estimated covariance matrix *S* to be fully estimable (of full rank), so that S^{-1} can be determined.

The mean vectors are compared using the two-sample Hotelling's T^2 statistic with homogeneous variance. The null hypothesis H_0 is tested with the Hotelling's T^2 statistic (Morrison, 1976, p. 137).

$$T^2 \equiv (\bar{F} - \bar{D})'[(\tfrac{1}{N_F} + \tfrac{1}{N_D})S]^{-1}(\bar{F} - \bar{D})$$

Under $H_{0,}$ the statistic $\dfrac{\nu - (p-1)}{p\nu} T^2 \sim F_{p,\nu-(p-1)}$ has a central *F* distribution with *p* and $\nu - (p - 1)$ degrees of freedom. Under the alternative hypothesis (that $\mu_D = \mu_F + \delta$), $\dfrac{\nu - (p-1)}{p\nu} T^2 \sim F'_{p,\nu-(p-1)}(\lambda)$, a noncentral *F* distribution with *p* and $\nu - (p - 1)$ degrees of freedom, and noncentrality parameter λ, where

$$\lambda = \delta' \left[\left(\frac{1}{N_D} + \frac{1}{N_F} \right) \Sigma \right]^{-1} \delta$$

This relation can be used to develop contours of differences between μ_F and μ_D that can be detected with high power. This power analysis can be used to determine whether differences between finished water and distribution water that are considered to be of toxicological importance can be detected with high power.

Detection of Outlying Processes—Mahalanobis D^2

A procedure similar to that discussed in the section on comparison of a new treatment process to a reference set of similar processes can be adapted to detect outliers among a reference set of M processes. Denote the reference set as $X_1, \ldots, X_M$ and let $\bar{X}$, *S* denote the mean vector and covariance matrix of the *X* terms. Let X_m denote the *m*th water sample, $m = 1, \ldots,$ M. For the *m*th process let X_m, denote the response and let $D^2{}_m$ denote the Mahalanobis D^2 statistic,

$$D_m^2 \equiv (X_m - \bar{X})' S^{-1} (X_m - \bar{X})$$

(Morrison, 1976, p. 235). For this analysis, all $M = 15$ finished water samples and distribution system water samples were treated as if they came from a single distribution.

The Mahalanobis D^2 is the square of the distance of each observation to the center of the distribution, accounting for the variances and the covariances in the data. It is similar to the Hotelling's T^2 statistic discussed earlier, except that *M* is taken to be sufficiently large that the $1/M$ term in the covariance expression for the Hotelling's T^2 statistic is considered sufficiently small to omit.

The overall Type 1 error level must be sufficiently conservative to allow for the number of anticipated water samples that are falsely declared to be potential outliers to be controlled to within an acceptable level. For example, if *M* water sample response vectors are to be screened, then the overall probability of falsely declaring a process input to be an outlier is to be controlled at an overall level α (e.g., 0.10), and if there are no outlying processes, then the significance level for each individual comparison might be set at α/M.

An observation X_m is classified as a potential outlier (i.e., potentially dissimilar) if D_m^2 is statistically significant. (A "potential outlier" is sometimes referred to in the statistical literature simply as an "outlier.")

The ordered D_m^2 terms are plotted in a chi-square probability plot. Treating the overall mean vector and covariance matrix as approximately known, the distances are distributed approximately as chi square with degrees of freedom equal to the dimension of the vector, *p* as *M* gets large. Under normality assumptions and in the absence of outliers, the plot should resemble a straight line with slope 1. Outlying values would lie above the curve through the majority of the chi square values. Figure 1 displays the Mahalanobis D^2 distances for the Schenck et al. (2008) data ($p = 7$). The reference line is the chi square distribution with *p* degrees of freedom. The plot conforms reasonably well to the order statistics from the chi-square distribution with $p = 7$ degrees of freedom.

A robust version of the outlier detection procedure involves removing one observation at a time when calculating the mean vector and the covariance matrix. This improves the sensitivity to detect outlying observations since the influence of an outlier on the robust mean and the robust covariance is removed. For each observation X_m, let $\bar{X}_{(-m)}$ *and* $S_{(-m)}$ represent the mean vector and the covariance matrix, respectively, of the $M - 1$ observations that remain after omitting X_m. Repeating this for each X_m, the robust version of the Mahalanobis D^2 statistic is

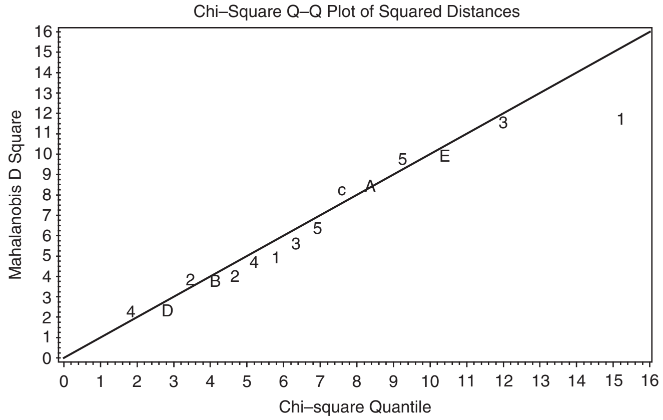

FIG. 1. Chi-Square Plot of Mahalanobis square distance. (Symbols A–E represent finished water and symbols 1–5 represent distribution water. Reference line is the chi square distribution cdf with $p = 7$ degrees of freedom).

calculated for $m = 1, \ldots, M$, where one observation is removed for each calculation. Denote the robust version of the already described distance function $D^2{}_m$ as $RD^2{}_{(-m)}$:

$$RD^2_{(-m)} = (X_m - \bar{X}_{(-m)})' \left[S_{(-m)} \right]^{-1} (X_m - \bar{X}_{(-m)})$$

An observation is classified as a potential outlier (i.e., potentially dissimilar) if $RD^2{}_{(-m)}$ lies above the curve through the majority of the values. Figure 2 displays a chi-square plot of the robust Mahalanobis D^2 distances for the Schenck et al. (2009) data ($p = 7$). The reference line is the chi-square distribution with p degrees of freedom. Two distribution water samples, from sources 1 and 3, appear to be potential outliers.

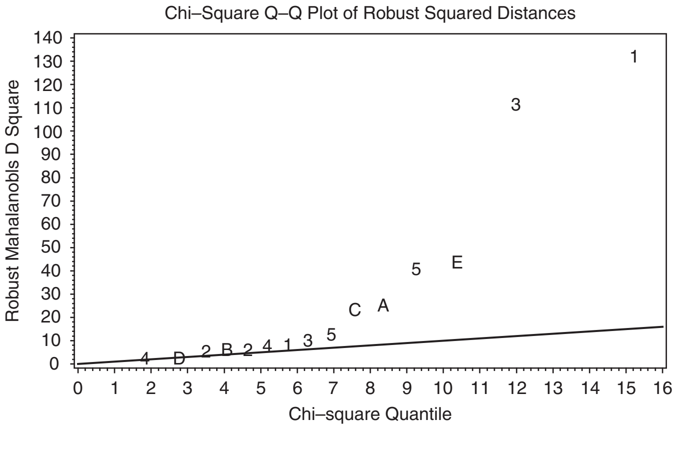

FIG. 2. Robust Chi-Square plot of Mahalanobis square distance. (Symbols A–E represent finished water and symbols 1–5 represent distribution water. Reference line is the chi square distribution cdf with $p = 7$ degrees of freedom).

Matched-Pairs Analysis

The matched-pairs analysis formulation incorporates the natural pairing that occurs when comparing the characteristics of the finished water with those of the distribution water from the same treatment plant. It treats the finished water and the distribution water samples collected from the same treatment plant as matched pairs. A matched-pairs analysis is a multivariate generalization of the two-way analysis of variance model, with blocks and treatments. The treatment plants represent the "blocks," and the finished water versus distribution water differences represents the "treatment effects." Let X_{ijk} denote the output characteristics of the water samples. Note that in this section the indices i and j have different meanings than in previous sections. Here $i = 1, 2$ indicate the finished water ($i = 1$) or the distribution water ($I = 2$). The index j corresponds to the treatment plants ($j = 1,. . .,5$). The index k corresponds to the replicate determinations for each water type and treatment plant combination. In the application to the Schenck et al. (2009) data there are $K_{1j} = 1$ replicate of finished water and $K_{2j} = 2$ replicates of distribution water for each plant j. Assume that (possibly after taking appropriate transformations) the random vector X_{ijk} is approximately distributed according to a multivariate normal distribution with common covariance matrix:

$$X_{ijk} \sim \text{MVN}(\mu_{ij}, \Sigma) \;\; i = 1, 2; j = 1, \ldots, 5; k = 1, \ldots, K_{ij}$$

Let $K_1 \equiv \Sigma_j K_{1j}$, $K_2 \equiv \Sigma_j K_{2j}$, and $N \equiv K_1 + K_2$. K_1 is the total number of finished water samples and K_2 is the total number of distribution water samples. The statistical model assumes constant covariance across the treatment plants and finished water versus distribution water. A large number of replicate determinations would be necessary within each treatment plant and water source combination to assess departures from the assumption of constant covariance.

The mean response vector μ_{ij} for water type i, plant j, can be expressed according to the model

$$\mu_{ij} = \mu + \alpha_i + \beta_j \qquad i = 1, 2;\ \mathrm{j} = 1, \ldots, 5$$

The α_i terms represent effects due to finished water ($i = 1$) and distribution water ($i = 2$) and the β_j terms represent the effects associated with the water treatment plants ($j = 1, \ldots, 5$). The parameters are subject to the constraints $\Sigma_i \alpha_i = 0$; $\Sigma_j \beta_j = 0$. These constraints imply that μ represents the average response across water types and treatment plants, $\mu + \alpha_i$ represents the average response for water type i averaged across treatment plants, and $\mu + \beta_j$ represents the average response for treatment plant j averaged across water types. In this formulation the treatment plants represent the totality of the treatment plants about which inferences are to be made based on the Schenck et al. data. Let r denote the number of independent parameters of the model (i.e., the "rank"). Thus $r = 1 + (2 - 1) + (5 - 1) = 6$.

The null hypothesis of equality of finished water and distribution water characteristics is to be tested based on this model:

$$\mathrm{H}_0\text{: } \alpha_1 = \alpha_2 = 0$$

versus

$$\mathrm{H}_1\text{: } \alpha_1 \neq \alpha_2$$

Morrison (1976, p. 182ff) discusses the analysis of this model using a randomized blocks analysis of variance approach. Let $\hat{\mu}_{ij}$ denote the estimate of μ_{ij} under the model just described. The error sums of squares and cross products matrix E is $\Sigma_{ij}\Sigma_k\,(X_{ijk} - \hat{\mu}_{ij})(X_{ijk} - \hat{\mu}_{ij})'$. The estimated error variance covariance matrix is $S \equiv \frac{1}{N-r}E$, where r (6) is the rank of the model. Let $\overline{X}_{i..}$ denote the average finished water sample response (i = 1) or distribution water response (i = 2) across all the treatment plants and replicate determinations within treatment plants. In the case of balanced sample sizes, the matched pairs analysis of variance statistic is proportional to

$$Q \equiv (\overline{X}_{1..} - \overline{X}_{2..})'\,\mathrm{E}^{-1}(\overline{X}_{1..} - \overline{X}_{2..})$$

This statistic resembles the Hotelling's T^2 statistic discussed earlier, except the error covariance matrix is estimated based on the residuals from the matched-pairs model after adjusting for differences among the treatment plants. Morrison (1976, pp. 178, 185) indicates that under $\mathrm{H}_{0,}$ the analysis of variance statistic can be related to a central F distribution with degrees of freedom p and $(N - r) - (p - 1)$, namely,

$$\frac{(N - \mathrm{r}) - (p - 1)}{p}\left[\frac{K_1 K_2}{K_1 + K_2}\right] Q \sim F_{p,\ (N-r)-(p-1)}$$

Under the alternative hypothesis, the statistic is proportional to a noncentral F distribution.

Linear Principal Components Analysis

Linear principal components analysis (referred to later simply as principal components analysis) is a data-analytic procedure that empirically can be used to represent most of the variation in a relatively high dimensional response vector, in a lower dimensional space. Each transformed variable is a linear combination of the original variables and is uncorrelated with the others. In this article, principal components analysis was carried out on the studentized residuals from the randomized blocks model fit. The results are used to visualize the data in a small number of dimensions, in a manner that reveals much of the structure and relationships in the data.

The first principal component direction is that direction, across all possible directions, that has the largest variance among the observations. The second principal component direction is that direction, orthogonal to the first principal component direction, that has the greatest variation among observations, across all possible directions orthogonal to the first principal component direction. This process continues until the remaining principal components account for the entire p-dimensional space, with successively less and less variance in each successive orthogonal direction. The successive orthogonal directions are referred to as the "principal component" directions. The sum of the variances across these principal components directions is equal to the sum of the variances across the originals variables. Principal components analysis is often a useful data analytic approach because it is often the case that the sum of the variances associated with the first few (e.g., two or three) principal components directions accounts for a substantial proportion of the total variation among all the original variables.

The principal components directions with the largest variances reflect the relationships among the responses such that the members of the sample have the greatest variation among them. Representing the data in the low dimensional space of the principal components with the largest variances will often reflect major relationships among the original responses. These directions are good to examine to identify characteristics of the underlying probability distributions and to look for departures from assumptions of a common distribution among the water sources. Often the data are examined graphically in the space of the principal components with the largest variances to determine whether the data (and associated water samples) cluster into subsets rather than arise from a single distribution.

It is often informative to examine the principal component directions with the smallest variances. These combinations of the original responses contain the smallest proportions of the variation in the reference set. Detection of outlying process vectors or of vectors corresponding to dissimilar mixtures may be easier in directions in which the reference set is most tightly bunched. Donnell et al. (1994), Jolliffe (2004, chap. 10), and references therein discuss applications of the principal components with the smallest variances. Jolliffe (2004, p. 235) states, "By examining the values of the last few PCs [principal components—PF[we may be able to detect observations that violate the correlation structure imposed by the bulk of the data, but that are not necessarily aberrant with respect to individual variables."

The method of principal components is mathematically based on the fact that the sample covariance matrix S of the original responses X can be factored into GLG', where the columns of G are orthonormal and L is a diagonal matrix with diagonal elements $/_1 \geq /_2 \geq \ldots \geq /_p$. The diagonal elements of L are referred to as the sample "characteristic roots." The columns of the orthonormal matrix G are referred to in mathematical terms as the sample "characteristic vectors" and in statistical terms as the "factor loadings" of the original responses on the principal components.

The principal components are studied as a data-analytic tool. The data are plotted in the directions associated with the largest characteristic roots in order to visualize the relationships among the treatment processes in lower dimensional space, and in the directions associated with the smallest characteristic roots to detect possible outlying responses or responses from another population.

EXAMPLES: APPLICATION TO SCHENCK ET AL. DATA—OUTPUT WATER CHARACTERISTICS

The methods discussed earlier are illustrated with data from the five water treatment plants that were monitored by Schenck et al. (2009). The output water samples from each treatment plant consisted of three samples—one finished water sample and two distribution water samples.

The five treatment plants, the month and the year during which they were monitored, the water source, and the chlorine treatment are summarized in Table 1. The water sources and the chlorine doses varied among the treatment plants. Each of the five treatment plants was monitored once at a single sampling point for raw input water and finished output water, and at two distribution system output water sites. Each plant used a chlorination process.

The 20 characteristics of the output finished water and distribution system water, listed in block 3 of Table 2, represent the set of outputs from the treatment plants. Because of the limited number of treatment plants and sampling times included in the data set, the following subset of seven output characteristics was selected to be used in the examples in this section:

- Total organic carbon (TOC) (mg/L).
- Total organic halogens (TOX) (μg/L).
- Mutagenic activity (revertants/L equivalent).
- Total trihalomethanes (TTHM) (μg/L).
- HAA6 (six haloacetic acids[a]) (μg/L).
- Percent brominated TTHM.
- Percent brominated HAA6.

Six of the seven output characteristics are chemical characteristics of the output water. Mutagenic activity is a toxicological characteristic. Schenck et al. (2009) reported that mutagenic activity is highly correlated with TOC and TOX.

The values of these seven output characteristics for the finished water samples and for both of the distribution system water samples from each treatment plant are displayed in Table 3. No transformations were carried out on the data prior to analysis. The univariate means and standard errors for the finished water and for the distribution water are displayed in Table 3 for each of the seven output characteristics and for each group. Except for TTHM, the differences in means between the finished water and the distribution system water samples lie well within a single standard error of the mean for each characteristic. For TTHM, the difference in means is 1.4 standard errors apart, which is not overly large, particularly for the largest difference among 7 characteristics. The result of the Hotelling's T^2 test for the two-sample comparison between the finished water (X) and the distribution system water (Y) is not statistically significant (p = .52). This statistic is distributed under the null hypothesis according to an F distribution with $p = 7$ and $(N_X - 1) + (N_Y - 1) - (p - 1) \equiv (5 - 1) + (10 - 1) - (7 - 1) = 7$ degrees of freedom.

Outlier Detection

Figure 1 displays a chi-square probability plot of the ordered Mahalanobis D^2 statistic $D_m^2 = (X_m - \bar{X})' S^{-1} (X_m - \bar{X})$ and Figure 2 displays a chi-square probability plot of the ordered robust version of the Mahalanobis D^2 statistic $RD_m^2 = (X_m - \bar{X}_{(-m)})' S_{(-m)}^{-1} (X_m - \bar{X}_{(-m)})$ (Morrison, 1976, p. 235) for each of the M = 15 sample responses from the overall mean response. For each m, the mean and covariance $\bar{X}_{(-m)}$ and $S_{(-m)}^{-1}$ are calculated, omitting the mth observation. These are outlier detection displays. For this analysis, all 15 finished water samples and the distribution system water samples were treated as coming from a single sample. Treating the overall mean vector and covariance matrix as approximately known, the distances are distributed approximately as chi square with degrees of freedom equal to the dimension of the vector, namely, p = 7. The letter plotting symbols correspond to the finished water samples from each treatment plant and the number plotting symbols correspond to the distribution water samples. Under normality assumptions and in the absence of outliers, the chi-square probability plot should follow the plotted chi-square cdf, which is a straight line with slope 1, going through the point (1, 1). Outlying values would lie above the plot of the remaining points.

In Figure 1 all the points but a groundwater distribution sample (1) lie close to the straight line corresponding to the chi-square reference distribution with p = 7 degrees of freedom. The groundwater distribution sample (1) has a smaller distance from the mean than would be predicated for the largest order statistic based on the reference distribution. Figure 1 displays no indications of outliers.

Figure 2 is based on the differences between each sample and the overall robust means, normalized by the robust covariance. The $RD^2{}_m$ values are ordered the same as in Figure 1. The points lie above the line corresponding to the chi-square distribution with p = 7 degrees of freedom. The points most removed from the chi-square distribution line correspond to finished water and distribution water samples from treatment plants 1, 3, and 5. In particular, two distribution system samples are apparent outliers, one from treatment plant 1 (groundwater source) and one from treatment plant 3 (river surface water source), in that they are removed from and well above the curve corresponding to the remaining 13 samples.

[a]The six haloacetic acids denoted as HAA6 are chloroacetic acid, bromoacetic acid, dichloroacetic acid, trichloroacetic acid, bromochloroacetic acid, and dibromoacetic acid.

TABLE 3
Selected Source Water Characteristics and Disinfection-By-Products in the Finished Water and Distribution System

Treatment plant	Sample	Total organic halogen (μg/L)	Total organic carbon (mg/L)	Mutagenic Activity	Total trihalomethanes (μg/L)	Six haloacetic acids (μg/L)	% of total trihalomethanes brominated	% of total six haloacetic acids brominated
	Distribution (1)	28.80	1.99	683.50	27.40	4.01	55.47	72.82
1: 1995	Distribution (1)	32.35	1.97	351.50	39.00	1.19	64.23	59.92
	Finished (A)	26.75	2.00	941.25	16.95	2.58	42.48	67.83
2: March, 1996	Distribution (2)	488.00	5.18	5555.50	82.50	109.30	0.00	0.02
	Distribution (2)	488.00	4.89	5379.50	98.70	116.10	0.00	0.07
	Finished (B)	448.00	4.99	5635.50	65.60	98.30	0.00	0.09
3: June, 1996	Distribution (3)	198.00	1.80	3993.67	90.00	66.00	0.00	0.08
	Distribution (3)	457.00	3.07	5254.33	178.20	126.10	0.00	0.00
	Finished (C)	231.00	1.71	3738.00	64.20	64.80	3.89	0.08
4: 1997	Distribution (4)	277.00	2.32	3614.67	99.20	78.00	0.00	0.00
	Distribution (4)	228.00	2.29	2773.00	115.50	35.00	0.00	0.00
	Finished (D)	265.50	2.38	3563.83	92.25	73.50	0.00	0.00
5: 1998	Distribution (5)	294.30	3.21	3453.00	67.80	0.00	0.00	0.00
	Distribution (5)	238.60	3.31	2977.33	77.50	0.00	0.00	0.00
	Finished (E)	322.80	3.52	5731.00	72.00	58.90	2.78	0.31
Mean (Standard	Distribution	273.01	3.00	3403.60	87.58	53.57	11.97	13.29
Error of the	(N = 10)	(53.04)	(0.38)	(573.96)	(13.23)	(16.44)	(8.01)	(8.90)
Mean)[a]	Finished	258.81	2.92	3921.92	62.20	59.62	9.83	13.66
	(N = 5)	(68.76)	(0.60)	(873.37)	(12.37)	(15.76)	(8.20)	(13.54)

[a] Hotelling's T^2 test indicates that overall sample effect is not significantly different between finished water and distribution water, $p = 0.52$.

The robust version of the Mahalanobis D^2 statistic is more sensitive in detecting "potential outliers" than is the originally proposed statistic. The question of whether the "potential outliers" are "true outliers" or are simply the most extreme values from a heavy tailed distribution needs to be decided by examination of the underlying laboratory records by the laboratory analysis staff to determine whether there were deviations from the study protocol in the manner in which the data were taken, whether the ambient conditions associated with those samples differed in some fundamental way with those of the other samples, whether there were clerical errors, or whether the deviations represent natural variation. Since the data values in Table 2 vary over multiple orders of magnitude, consideration could be given to determine whether the use of preliminary transformations, such as log transformations, might bring these two samples into better agreement with the others.

Although there is no absolute right or wrong decision approach with how to deal with outlying values, the philosophy is adopted that unless the outlying values are known to have resulted from errors, from protocol deviations, or from unique ambient conditions that differ from the remainder of the data, the values are considered to be valid data, to represent natural variation, and they are retained in the analyses. There is no external information to suggest that the extreme samples in Figure 2 are associated with measurement errors of protocol deviations. The subsequent analyses are carried out including these outlying water samples.

Matched-Pairs Analysis

Analyses were performed that accounted for potential differences among treatment plants. A two-way multivariate analysis of variance model, which included effects of sample type (finished water or distribution system water) and treatment plant, was fitted to the data. This is effectively a matched-pairs analysis with paired responses within each treatment plant. The two-way multivariate analysis of variance model was fitted to the data using the GLM procedure in the SAS System (SAS Institute, Inc., 2004). The test for finished water versus distribution water effect is directly analogous to the Hotelling T^2 test considered previously, except the error sums of squares and cross-products matrix E is based on the residual vectors from the two-way randomized blocks model, after accounting for the effects of both plants and sample type. For purposes of the matched-pairs analysis, X_{ijk} corresponds to the response vector from replicate k of sample type i, $i = 1, 2$, and treatment plant j, $j = 1, \ldots, 5$. $\bar{X}_{1..}$ corresponds to the average response of finished water samples and $\bar{X}_{2..}$ corresponds to the average response of distribution water samples. The statistic

$$\frac{\upsilon-(p-1)}{p}(\bar{X}_{1..}-\bar{X}_{2..})'\left[\left(\frac{1}{K_1}+\frac{1}{K_2}\right)E\right]^{-1}(\bar{X}_{1..}-\bar{X}_{2..})$$

has an F distribution with degrees of freedom $p = 7$ and $\nu - (p - 1) = (K_1 + K_2 - 1) - (I - 1) - (J - 1) - (p - 1) \equiv (5 + 10 - 1) - 1 - 4 - (7 - 1) = 3$. K_1 and K_2 were defined in the Matched-Pairs Analysis subsection of the Statistical Methods section. Under the null hypothesis, the test for finished water versus distribution water effect is distributed as an F distribution with 7 and 3 degrees of freedom. Under the alternative hypothesis the test statistic has a noncentral F distribution. The error sum of squares and cross products matrix E includes the interaction of sample type and treatment plant pooled with replicate variation. There is no evidence of a statistically significant sample type effect ($p = .55$).

TABLE 4
Analysis Results from the Two-Way Multivariate Analysis of Variance (MANOVA) on the Selected Source Water Characteristics and Disinfection-By-Products in the Finished Water and Distribution System

	Samples	Treatment plants
MANOVA Test for no overall effects	p-value = 0.55	p-value = 0.0042

The test for treatment plant effect is an analysis of variance test to compare the five treatment plant mean vectors. Under the null hypothesis, an analysis of variance test (Hotelling–Lawley trace—an extension of the Hotelling's T^2 test for the comparison of more than two groups) for treatment plant effects has a distribution that is approximated by an F distribution with 28 and 2.57 degrees of freedom. There is strong evidence of a difference among treatment plants ($p = .004$). This difference is in part due to differences between groundwater and surface water sources, as well as differences among treatment plants, as seen in Figure 4, to be discussed later. The results of the analysis of variance tests are displayed in Table 4.

Principal Components of Residuals From Matched-Pairs Analysis

A principal components analysis was performed on the studentized residuals from the two-way multivariate analysis model that includes effects due to treatment plant and sample type. Studentized residuals from the multivariate analysis of variance model were determined by leaving out one observation at a time, determining the residuals from the two-way analysis of variance model fitted to the remaining data, and normalizing the residuals by their standard errors in a component-wise basis. The studentized residuals reflect the correlation structure among the responses in the data.

The results are displayed in Table 5. The "loadings" of the original variables on the first principal component are denoted as "Char Vect 1." The loadings result in an essentially unweighted average of four output DBP constituents and mutagenic activity. They give very small weight to the percent bromination responses. Schenck et al. (2008) noted that the output DBP constituents in the first principal component are

TABLE 5

Loadings of the Original Variables on the Principal Components (Characteristic Vectors) and Variances of the Principal Components (Characteristic Roots) Based on the Principal Components Analysis of the Studentized Residuals in the Matched Pairs Analysis After Adjusting for Water Source and Treatment Plant Effects

	Char vect 1	Char vect 2	Char vect 3	Char vect 4	Char vect 5	Char vect 6	Char vect 7
Total organic halogen (µg/L)	0.4568	0.0215	–0.1162	0.1966	–0.7042	–0.1155	0.4790
Total organic carbon (mg/L)	0.4543	–0.0357	–0.2383	0.2853	0.1918	–0.6255	–0.4755
Mutagenic activity	0.4335	0.0042	0.4425	–0.4642	0.4031	–0.2574	0.4148
Total trihalomethanes (µg/L)	0.4366	0.0246	–0.3805	0.2801	0.4547	0.5870	0.1853
Six haloacetic acids (µg/L)	0.4524	0.0732	0.3217	–0.2844	–0.3011	0.4213	–0.5809
% of total trihalomethanes brominated	–0.0396	0.7413	0.4488	0.4908	0.0770	–0.0116	0.0225
% of six haloacetic acids brominated	0.0149	–0.6654	0.5332	0.5140	0.0390	0.0833	0.0115
Char roots and percent of total variance explained by the principal components	4.53 (60.7%)	1.59 (21.3%)	0.75 (10.0%)	0.37 (5.0%)	0.11 (1.5%)	0.07 (1.0%)	0.045 (0.6%)

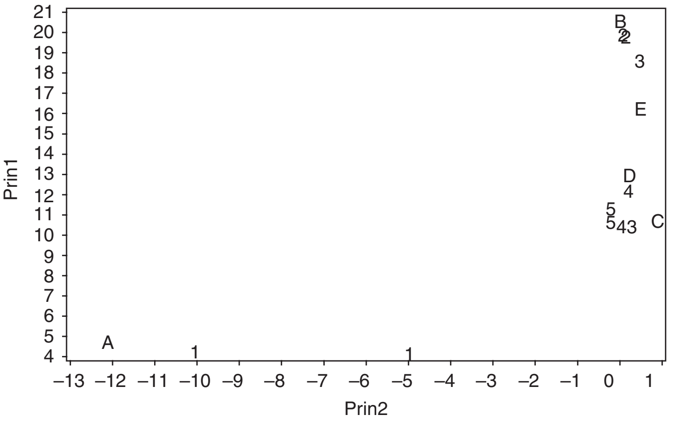

FIG. 3. First principal component vs. second principal component. Matched pairs analysis, adjusting for variability among treatment plants.

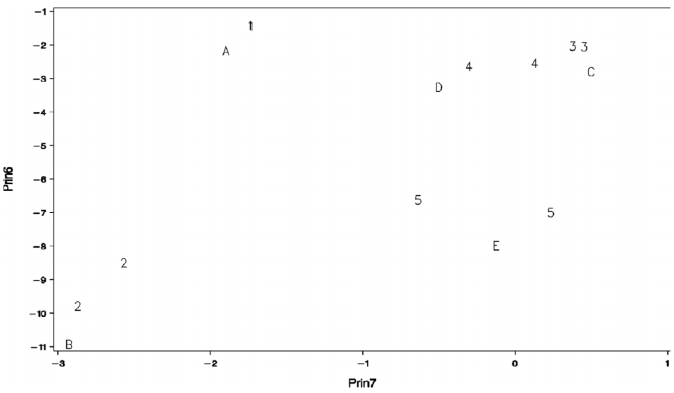

FIG. 4. Sixth principal component vs. seventh principal component. Matched pairs analysis, adjusting for variability among treatment plants.

positively correlated with mutagenic activity. The "loadings" of the original variables on the second principal component ("Char Vect 2") are essentially the difference between percent of TTHM that is brominated and percent of HHA[6] that is brominated. The output characteristics that dominate the first principal component have little weight in this component.

For each treatment plant and output water sample type the normalized mean scores (i.e., the normalized mean responses expressed in the principal components coordinates) were added back to the studentized residual scores. Figure 3 and Figure 4 display the re-centered studentized residual scores from each of the treatment plants in the principal components space. These figures display the relationships among the treatment plants and the water sample types in a series of statistically independent transformed variables, reflecting the greatest variation among the water samples and the least variation. Figure 3 displays the first and second principal components (i.e., those that explain the greatest portion of the variation), while Figure 4 displays the sixth and seventh principal components (i.e., those with respect to which the data have the least variation). In these figures the letter plotting symbols (A, B, C, D, and E) correspond to the finished water samples and the number plotting symbols (1, 2, 3, 4, and 5) correspond to the distribution water samples. Figure 3 clearly shows that the output characteristics associated with samples "A" and "1" (from the treatment plant with a groundwater source) are different from those of the other samples (from the treatment plants with the surface water sources) with respect to the first two principal components. Table 3 shows that the groundwater finished water and distribution water samples are highly brominated and have relatively low TOX, TTHM, and HAA6 concentrations, as well as relatively low mutagenic activity. The primary role of the first two principal components is to separate groundwater source treatment plant samples from the others. In Figure 4, based on the principal components with the least variation, the five treatment plants are clearly separated, which is in agreement with the finding of a highly significant treatment plant effect. Within each plant, the finished water samples and the distribution system water samples cluster together, which is in agreement with the finding of a nonsignificant sample type effect.

In summary, the graphical displays of the largest and smallest principal components of the Schenck et al. (2009) data indicate that the characteristics of the output from the treatment plant that uses a groundwater source differ considerably from the outputs from the four treatment plants that use surface water sources. The characteristics of the finished water and the distribution water from the same treatment plant are in agreement. This provides a visualization of the results of the statistical tests that indicated that there is no statistical evidence of differences in output characteristics between the finished water and the distribution system water, but there is statistical indication that the output characteristics differ among the treatment plants. These differences exist primarily between the groundwater source plant and the surface water source plants, but also exist among the four surface water source plants. Figures 3 and 4 raise a conjecture that treatment plants C, D, and possibly E are more similar to one another than to treatment plants A and B.

DISCUSSION

This article has reviewed and discussed several multivariate statistical analyses and display techniques useful for assessment of sufficient similarity of complex environmental mixtures. These methods were applied to (1) the comparison of inputs to and output water from treatment plants, or from different stages within the same treatment plant, (2) the determination of whether there is statistical evidence of heterogeneity, and if so, (3) the nature of the heterogeneity. These procedures illustrate several approaches for displaying and analyzing high-dimensional data that naturally arise in the comparisons of alternative water sources, the disinfection treatment processes that are applied to them, and the characteristics of the DBP in the output for the treatment plants. The methods discussed are based on the physical, chemical, and toxicological characteristics of the water sources. Feder et al. (2009) discuss the use of robust, nonparametric bootstrap methods to test the homogeneity of treatment plant water characteristics, as an alternative to the parametric normal theory analysis of variance procedures discussed in this article.

The analyses discussed in this article pertain to the statistical significance of differences among water samples. Statistically significant differences between water samples are those that are too large to have occurred just due to chance. This differs from the notion of toxicological significance, which generally means that the observed differences between water samples are sufficient to exert important health implications, but also, in some studies, a rare effect may be present in only one or a few test subjects but still be considered toxicologically significant. In most cases, however, a study is said to be "adequately powered" if differences that are sufficient to be toxicologically significant are also large enough to be statistically significant. This implies that differences sufficient to have health implications are likely to be detectable from the data.

The sensitivity of statistical analyses is impacted in this article by a number of considerations:

1. The selection of process input and output variables, so that the statistical analyses are sensitive to toxicologically significant effects.
2. The incorporation of valid estimates of variability within the data to construct valid and efficient statistical estimates and tests. True replicate determinations need to be built into the routine monitoring procedures to provide a basis for determining the variability in the responses. True replication needs to reflect all important sources of variation, in contrast to multiple measurements on the same samples, which may reflect only the analytical portion of the variation.
3. The deletion of water samples corresponding to potentially outlying responses should be confirmed on chemical and toxicological grounds, if possible.
4. The Schenck et al. (2009) data set is useful for illustrating several of the suggested analysis procedures. However, the data represent a relatively small set of treatment plants, not necessarily similar, each monitored at a single time point, and without true replicate measurements on which to base variability estimates.
5. Principal components analysis was demonstrated to be a useful data-analytic tool to help visualize and understand relationships portrayed in the data.
6. The inference procedures discussed in this article are based on process input and output chemical compositional characteristics and a calculated toxicological characteristic, mutagenic activity. Chemical compositional characteristics of the water samples are useful insofar as they are associated with toxicological characteristics.
7. As an extension of the methods discussed in this article, the analyses discussed could be performed as weighted analyses, with weights proportional to the relative toxicities of each mixture component. Such weights might be cancer potency factors, benchmark doses for noncancer endpoints, or relative concentrations compared to a no-effect or low-effect level. A logical consequence of incorporating toxicologically based weights into the analysis is that the weights, and therefore the analysis results, will be dependent on the particular health outcome being considered. Thus, mixtures may be "similar" with respect to certain health outcomes but not others.

Other statistical methods have been proposed based on components of the whole mixture as well as the whole mixture. Gennings et al. (1997) and Teuschler at al. (2000) consider statistical design and analysis methods for detecting departures from additive component effects in the toxicity of mixtures. This has important implications concerning the extent to which the toxicity of complex mixtures can be studied based on tests with individual mixture components or with defined mixtures involving combinations of components. Stork et al. (2008) present statistical methods for addressing sufficient similarity in dose-response for various mixtures of 11 chemicals containing the same components but with varying ratios.

REFERENCES

Box, G. E. P. 1949. A general distribution theory for a class of likelihood criteria. *Biometrika* 36:317–346.

Bull, R. J., Krasner, S. W., and Daniel, P. A. 2001. *The health effects and occurrence of disinfection by-products, Study 448*. Denver, CO: American Water Works Association and AWWA Research Foundation.

Bull, R. J., Teuschler, L., and Rice, G. 2009a. Determinants of whether or not mixtures of disinfection by-products are similar. *J. Toxicol. Environ. Health A*. 72:437–460.

Bull, R. J., Rice, G., Teuschler, L., and Feder, P. 2009b. Chemical measures of similarity among disinfection by-product mixtures. *J. Toxicol. Environ. Health A*. 72:482–493.

Donnell, D. J., Buja, A., and Stuetzle, W. 1994. Analysis of additive dependencies and concurvities using smallest additive principal components. *Ann. Stat*. 22:1635–1673.

Feder, P. I., Ma, Z., Bull, R. J., Teuschler, L. K., and Rice, G. 2009. Evaluating sufficient similarity for disinfection by-product (DBP) similar mixtures with bootstrap hypothesis test procedures. *J. Toxicol. Environ. Health A*. 72:494–504.

Gennings, C., Schwartz, P., Carter, W. H., Jr., and Simmons, J. E. 1997. Detection of departures from additivity in mixtures of many chemicals with a threshold model. *J. Agric. Biol. Environ. Stat*. 2:198–211.

Jolliffe, I. T. 2004. *Principal component analysis*, 2nd ed. New York: Springer Science and Business Media, LLC.

Morrison, D. F. 1976. *Multivariate statistical methods*, 2nd ed. New York: McGraw-Hill. (4th Ed., 2002, Brooks/Cole.)

Rice, G. E., Teuschler, L. K., Bull, R. J., Simmons, J. E., and Feder, P. I. 2009. Evaluating the similarity of complex drinking-water disinfection by product mixtures: Overview of the issues. *J. Toxicol. Environ. Health A*. 72:429–436.

SAS Institute, Inc. 2004 *SAS/STAT® 9.1 user's guide*. Cary, NC: SAS Institute, Inc.

Schenck, K. M., Sivagnasen, M., Rice, G. E. 2009. Correlations of water quality parameters with mutagenicity of chlorinated drinking-water samples. *J. Toxicol. Environ. Health A*. 72:461–467.

Stork, L. G., Gennings, C., Carter, W. H., Jr., Teuschler, L., Carney, E. W. 2008. Empirical evaluation of sufficient similarity in dose-response for environmental risk assessment of chemical mixtures. *J. Agric. Biol. Environ. Stat*.

Teuschler, L. K., Gennings, C., Stiteler, W. M., Hertzberg, R. C., Colman, J. T., Thiyagarajah, A., Lipscombe, J. C., Hartley, W. R., and Simmons, J. E. 2000. A multiple purpose design approach to the evaluation of risks from complex mixtures of disinfection by products (DBPs). *Drug Chem. Toxicol*. 23:307–321.

U.S. Environmental Protection Agency. 1986. *Guidelines for the health risk assessment of complex mixtures*. EPA/630/R-98/002. Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. http://www.epa.gov/ncea/raf/pdfs/chem_mix/chemmix_1986.pdf.

U.S. Environmental Protection Agency. 2000. *Supplementary guidance for conducting health risk assessment of chemical mixtures*. EPA/630/R-00/002. Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. http://www.epa.gov/ncea/raf/pdfs/chem_mix/chem_mix_08_2001.pdf.